Is the Notion of Validity Valid in HCI Practice?

Gitte Lindgaard

Human Oriented Technology Lab (HOT Lab), Carleton University, Ottawa, Canada.

Abstract. Much attention has been paid in the recent literature to the notions of validity, thoroughness, and effectiveness of different Usability Evaluation Methods (UEMs). Calculation of these makes sense if a study aims to compare UEMs, but not, it is argued here, if a study aims to evaluate a given application. Illustrated by a case study employing different UEMs, it is argued here that for practitioners, UEMs serve to supplement, rather than compete against, each other. The study shows also that it is not possible to calculate validity, thoroughness, and effectiveness in actual usability studies.

Keywords. Usability evaluation method, Heuristic evaluation, User test, Validity.

1. Introduction

Studies comparing different Usability Evaluation Methods (UEMs) have appeared in the literature for over a decade (Jeffries, Miller, Wharton and Uyeda, 1992; Nielsen and Molich, 1990), usually aiming to assess the relative efficacy of competing UEMs in order, eventually, to proclaim a "winner". Many of these studies rely on simple counts of raw usability "problems", even where there are multiple instances of the same problem, and without consideration of the relative seriousness of different problems. There is no agreed-upon standard for comparing methods, so in an effort to establish such standards, researchers have borrowed concepts from signal detection theory (Swets, 1954). Thus, a "hit" is a problem identified by a method that turns out actually to be problematic for users; a "miss" is the inability of a method to identify a known problem; a "false positive" is the flagging of a problem by one method that turns out not to be problematic for users in other, usually in user performance, tests, and a "false negative" is the dismissal of a problem that turns out to be problematic. (Sears, 1997; Hartson, Andre and Williges, 2003). Using these concepts, Sears (1997) provides equations for assessing the thoroughness, validity, and effectiveness of UEMs.

In most of these studies, the outcomes of analytic, or inspection methods, have been compared with performance-based methods. Among the former, Heuristic Evaluation (HE), and Cognitive Walkthroughs (CW) have received most attention (Cuomo and Bowen, 1994; Sears, 1997; Lavery, Cockton and Atkinson, 1997), and most prominent among the latter is Gray and Salzman's (1998) detailed review of a sample of five popular HCI studies that, among the UEMs these used, employed experiments.

Gray and Salzman (1998) reviewed how the studies were designed and conducted as well as the authors' interpretation of their data. They were specifically concerned with the observation that authors' claims invariably went well beyond what the data could support. Pessimistically, they concluded that all suffered from a lack of validity in some form. One problem with this review, noted in several invited comments (Olson and Moran, 1998), is the neglect to consider that all the studies were performed in the context of actual usability work rather than in research laboratories. The control afforded laboratory studies is absent in practice. As a researcher seeking specifically to establish the relative "value" of different methods your main concern is precisely with the factors highlighted by Gray and Salzman; as a practitioner, however, your task is to deliver results, in a short time and with limited resources, that are of value to the project team, to the product, and to the business. Thus, research designed to assess the relative efficacy of different UEMs must uncover *all* usability problems in a given application; in practice, the goal is to identify as many problems as possible, knowing that there are still some left in the product, which may then be addressed in a later version (Donoghue, 2002). Most comparison studies ignore this important difference between research and work contexts.

Usually, practitioners rely on triangulation, the "drawing on experience from different sources to assess, through reinforcement and correspondence with one's own experience, whether one is on the right track" (McClelland, 1998, p. 287). In practice, different UEMs are thus employed to provide *supplementary* information rather than *competing* with one another. For example, experiments usually test a sample of user tasks, not the entire population of tasks. In order to select to-be-tested tasks wisely, one needs to know something about users' context, about other tools they use in conjunction with the application, the range of typical tasks and those that are central for users and critical for the business as well as about those that may currently cause problems or be too cumbersome for users (Lindgaard, 1994). Such knowledge does not, and is not intended to, come from experiments, but is likely to emerge from observations of users, interviews with stakeholders, and so forth; hence the application of different UEMs in HCI practice.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
00 JUN 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Is the Notion of Validity Valid in HCI Practice?	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Human Oriented Technology Lab (HOT Lab), Carleton University, Ottawa, Canada.**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001766, Work with Computing Systems 2004 (Proceedings of the 7th International Conference)., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**5**	

In this paper we argue provocatively that the notions of UEM validity, thoroughness, and effectiveness serve no useful purpose in practical usability evaluations, and that it can be quite legitimate for practitioners to draw conclusions that go well beyond what the usability data can substantiate. A case study is presented to support this argument.

2. The Website Investigated

A Danish local government web site was evaluated. Local government plays a significant role in the day-to-day life of ordinary citizens in Denmark; most social and other services (schools, kindergartens, unemployment benefits, home nursing, elderly care, etc.) are provided at this level of government. The home page of the website is shown in Figure 1 below.

Figure 1. The homepage of the web site investigated

3. Method

Three UEMs were employed in the evaluation: (1) interviews with the site stakeholders, (2) a HE, and (3) a user test. Unstructured interviews with employees of the relevant government office yielded information about the purpose of the site, the users in the district, the resources available to build and maintain the site and perceived problems with the content or the usability of the site. The HE was conducted by a single evaluator in two sessions; one lasting 2 hours 45 minutes, and one taking one hour. The findings of the HE were used, together with the outcomes of the interviews, to design a set of typical user tasks, which were then tested on a sample of eight users. Nielsen and Landauer (1993) assert that this should be enough to uncover 80% of the usability problems, although this is based on the assumption that each subject finds nearly one half of the problems known to exist in the application. This assumption is not always met (Lewis, 1994). Because the evaluation was performed in Australia and the web site was intended for Danish citizens, it was not possible to interview or test users in the relevant district. However, participants in the user test all spoke fluently Danish. Some were Danish citizens.

4. Interviews with Site Providers

Five people employed by the relevant government office took part in unstructured interviews taking up to one hour each. The IT manager had designed the website and was responsible for keeping it up to date. One of his assistants checked the site every day and responded to users' comments and questions or forwarded these to the relevant experts. Another person who was in charge of the taxation office was responsible for providing timely tax-related information to citizens and to deal with users' online tax questions. The fourth person was the district nurse who provided information about her services, office hours, contact details, and so forth. The final person interviewed acted as a link to the two nursing homes in the district. Her role was to deal with residents and staff at the nursing homes as well as answer online queries and provide information about services at the nursing homes to the public. For example, one home opened its cafeteria to the public on certain days of the week as well as providing exercise classes for seniors. The interviews pointed to incomplete content and to certain difficulties users apparently experienced when attempting to locate information.

5. Heuristic Evaluation

A HE was performed to understand where and why users may encounter stumbling blocks, and to use these to design tasks exposing a maximum number of perceived problems in the subsequent user test. One highly experienced evaluator performed the evaluation. Nielsen's (1993) published list of 10 heuristics was used to guide the evaluation.

Problems identified were counted in two ways. First, unique problems, also referred to here as "problem types", represented specific problem categories. For example, the HE showed that different search terms were required to locate same-type services in the district (e.g. kindergardens), and it was difficult visually to distinguish buttons from non-clickable images. In terms of the heuristics, same-type services requiring different search terms to be located is a "standards and consistency" problem, whereas buttons that do not indicate visually that they are clickable are "visual feedback" problems. In another example of a "visual feedback" problem different-coloured dots on a map of the district indicated the location of various services, but there was no mouse-over text or legend to communicate this to the user. Thus, different problem types were subsumed in a given heuristic. Some 56 unique problem types were identified.

The second way problems were counted was by including all instances in which each problem type occurred. Thus, for example, two instances of "dead links" were noted as a single problem type but as two problem tokens. Scoring problems in this way ensures that none will be forgotten when the site is revised, and the number of violations of any heuristic helps raise designers' awareness of those aspects of the interface that may need special attention in later revisions and in user tests. Another issue is that the heuristics were vaguely defined and are not mutually exclusive. Therefore, a given problem may equally violate two heuristics at the same time. These instances of heuristic violations are referred to here as "problems tokens"; up to a total of 154 tokens were

identified. The distribution of these is shown in Table 1 below. As the Table shows, the most frequently occurring tokens were related to visual feedback (n = 36), standards and consistency (n = 33), flexibility (n = 20), and visibility of system status (n=18).

Table 1. Number of violations for each heuristic

Rule	Heuristic	N
1	Match with user's task	10
2	**Standards and consistency**	**33**
3	**Visibility of system status**	**18**
4	Error prevention	8
5	Error recovery	2
6	User control and feedback	13
7	**Visual feedback**	**36**
8	Aesthetic and minimalist design	8
9	Recognition rather than recall	6
10	**Flexibility and efficiency of use**	**20**
	Total	**154**

The degree to which the 56 problem types were perceived to obstruct users' efforts to retrieve information effectively and efficiently was estimated by rating the severity of each along a 5-point scale ranging from 'Show stopper' to 'Minor irritation'. There were four 'show stoppers', 10 'severe hindrances', and 23 'hindrances'. The remaining nineteen problems were minor irritations. Thus, 25% (n = 14) of the unique problems were very serious or critical.

6. User Test

Seven tasks were designed to expose as many of the most severe problems and of the heuristic violations as possible. In terms of severity, the tasks included some 24 problem types, namely all four 'show stoppers', seven of the 10 'severe hindrances', and 10 of the less severe problems. In terms of problem tokens, some 16 of the 36 'Visual Feedback' tokens were exposed as were 13 of the 33 'Standards and consistency' tokens, seven of the 20 'Flexibility and efficiency of use' tokens, and five of the 18 'Visibility of system status' tokens. Four of the tasks were labeled 'simple' because the target information could be reached via more than one navigation path. The remaining three tasks were labeled 'complex', as the relevant information could only be retrieved via a single navigation path.

The number of clicks necessary to complete each task had been calculated a priori for all possible navigation strategies, and the minimum number of clicks was taken as the optimum against which performance was measured. Navigation strategies were analyzed, but as they are irrelevant for the argument made in this paper, they are not discussed further here. For more detail, see Lindgaard (1999).

6.1. Procedure

All subjects attempted the seven tasks which were given in the same order. Tasks were presented one at a time on a card to prevent subjects from reading ahead and attempting tasks in a different order. The same order was used because it was predicted that some of the tasks would prove too difficult to complete. Thus, to motivate subjects and ensure they would experience some level of success, four simple tasks were presented first [Tasks 1- 4], followed by three complex tasks [Tasks 5 – 7). Simple tasks could be accomplished by using any one of three navigation strategies; complex tasks could only be accomplished by one strategy. Subjects were instructed to work through the tasks at a comfortable pace and to refrain from asking questions of the experimenter once the session was underway. Although they were allowed to give up if an answer could not be found, this was not explicitly mentioned. The experimenter was seated just behind the subject during the session, taking notes to determine the search strategy and number of clicks was recorded electronically. Task performance was timed from the moment the subject had read the task until the answer had been found, or until the subject announced they were giving up.

6.2. Results

On average, subjects completed 4.86 (69%) of the seven tasks correctly. No one completed all seven tasks, but all subjects attempted all tasks. The number of correct answers given to each task is shown in Table 2.

Table 2. Number of Correct Answers for the Seven Tasks (N=8 subjects)

Task type	Task	Number correct
Simple	1	7
Simple	2	8
Simple	3	8
Simple	4	8
Complex	5	0
Complex	6	5
Complex	7	4

Clearly, performance reflected the division of tasks into simple and complex, as tasks 1-4 yielded correct answers in 31 of 32 possible hits. By contrast, tasks 5-7 resulted in correct answers in only nine out of a possible 24 hits. Thus, while everyone, with the exception of one subject on one task, completed the four simple tasks correctly, no one completed all the complex tasks. Task-completion times showed that complex tasks took about 30% longer than easy ones. Unsuccessful complex tasks took around 50% longer than successful ones, suggesting that subjects were taking the tasks seriously. Even when subjects were successful, they invariably used twice the number of clicks necessary to locate the answer. Thus, this measure did not reflect task difficulty. When subjects were unsuccessful, the number of clicks was 300% more than necessary, suggesting that they did not give up prematurely. Since the objective of the user test was to confirm that at least those problems labeled serious in the HE actually comprised user stumbling blocks, and as the purpose of the website is to serve all citizens, it was not of interest to apply inferential statistics.

6.3. Hits and false positives

In order to illustrate the argument to be made here, only relevant results are shown and only for the heuristics that were violated most often. Table 3 below shows the "hits", the "false positives", the total number of perceived problems exposed in the user test, and the total number of perceived problems identified in the HE. Since a user test is specifically designed to expose only a sample of problems and possible user tasks, the "total tested" problem column shows only the number of tokens for each of the four most frequently violated heuristics. The "total found" column shows the total number of tokens identified for the relevant heuristic by the HE.

Table 3. Number of hits, false positives, problems tested and problems found by the HE

Heuristic	Confirmed "hits"	Not confirmed "false positives"	Total tested	Total found
Visual feedback	12	4	16	36
Standards and consistency	8	5	13	33
Flexibility and efficiency of use	5	2	7	20
Visibility of system status	5	0	5	18
Total	**30**	**11**	**41**	**107**

The sample of "confirmed" problems included all four "show stoppers" and six of the seven "severe hindrances". The rest were less severe problems and some that occurred more than once in the tasks. When a given token was encountered in more than one task and was confirmed in one but not in another, it was entered in both the "confirmed" and the "not confirmed" columns.

6. Discussion

For the purpose of our argument, we may regard the "total number of problems found" by the HE as the "problem set". The "Confirmed" or "hits" are those that were identified by the HE and that also caused problems in the user test. These are thus the "real" problems. The "Not confirmed" column contains those that were predicted to be problematic from the HE but which were not confirmed by the user test. These represent the "false positives". There were no "misses" – problems found in the user test that had not been identified in the HE, and no "false negatives" – issues rejected by the HE but found to be problematic in the user test. Since the calculation of "thoroughness", "validity", and "effectiveness" all take into account the total number of problems in the application, none of these can be estimated from the above data because only a subset of problems was tested. This is shown in the column labeled "total tested", whereas the "Total found" column shows the number of problems identified by the HE. Yet, even if it were possible to calculate thoroughness, validity and effectiveness, it is questionable that anything useful would be achieved by doing so. Recall that the numbers in the Table refer to tokens, that is, individual occurrences of a smaller sample of problem types. Some of the tokens from each problem type were confirmed by the user test and others were not. Thus, for example, lack of visual feedback upon a given action caused a problem in one user task but not in another. Therefore, the degree to which a given problem type comprises a stumbling block for users appears to be context-dependent. Likewise, certain tokens representing the same problem type turned out to be stumbling blocks for users in one task but not in another. Because any problem found to be a stumbling block in any of the tasks appears in the "confirmed" column even if it was not problematic on every occasion on which it was encountered, the status of these cannot be determined because of this context-dependency. In addition, we cannot be confident that every problem type and every token has been identified by the HE. Therefore, there is no certainty that our problem set is complete. A method may thus result in a "thoroughness" index of, say, .94 (out of a maximum of 1.0), but there is no guarantee that the evaluation itself is "thorough". Similarly, the very fact that only a subset of problems and of possible user tasks is tested renders the notion of "thoroughness" impossible to assess and meaningless to calculate in the context of an actual evaluation. Likewise, to produce some figure representing the "validity" or "effectiveness" of one or more of the methods employed is just as meaningless and adds no value to the outcome; the purpose is to evaluate the application, not to assess the effectiveness of one or more UEMs. None of these dimensions can help to determine which problems should be fixed and which ones may wait till the next version or even which ones could safely be ignored.

The website would not be "fixed" if we were to change all the tokens found to be problematic in the user test because the problem types pervade the entire product. To fix the website, a complete revision is required instead. However, the above results should be applied to improve the site. Visual feedback and visibility of system status are both aspects of the visual UI design; flexibility relies on decisions about navigation structures and UI architecture; issues of consistency and standards relate to interface style compliance. Since these four heuristics were violated most frequently, the website would probably improve dramatically by thoroughly revising those aspects of the design. However, the observation that a given token was a stumbling block in one but not in another task is problematic, especially when the problem type occurs in multiple points in the website. It is unclear from the above findings when such a problem should be solved and when it should not, since not all instances were tested.

The conclusion that the website should be completely revised along the dimensions mentioned above is only partly supported by the test/evaluation data. The interviews pointed to certain problems that were not found in either the HE or the user test. Some of these were problems of omission – missing information, and some were problems of commission. Without thorough knowledge of all the users' tasks, context, and information requirements, neither the HE nor the user test can identify omissions. Thus, our recommendation to add the areas of information identified as missing by the interviews but not by other UEMs violates Gray and Salzman's (1998) assumption that good science equals good practice, that validity equates value, and that conclusions should never go beyond the story one's data can substantiate. That is, of course, unless interview data can be taken as "data" in the

sense Gray and Salzman discuss. The heuristic evaluation yielded information about specific UI dimensions to be revised, and it provided a number of placeholders, issues that can serve as benchmarks when testing the revised edition. So, despite the probable incompleteness of the problem set, the testing of only a subset of possible user tasks, and the uncertain status of some of the data, a revision is likely to result in a better end product even though it is only partially supported by the data, the validity of which we cannot even ascertain. We therefore argue strongly that the notions of validity, thoroughness, and effectiveness are of little value to the HCI practitioner.

7. References

Cuomo, D.L., & Bowen, C.D. (1994). Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting with Computers*, *6* (1), 86-108.

Donoghue, K. (2002). *Built for use: driving profitability through the user experience.* New York: McGraw-Hill.

Gray, W.D., & Salzman, M.C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, *13*, 203-262.

Hartson, H.R., Andre, T.S., & Williges, R.C. (2003). Criteria for evaluating usability methods. *International Journal of Human-Computer Interaction*, *13 (4),* 373-410.

Jeffries, R., Miller, J.R., Wharton, C., & Uyeda, K.M. (1992). User interface evaluation in the real world: A comparison of four techniques. *Proceedings CHI Conference on Human Factors in Computing Systems* (pp. 119-124). New York: ACM Press.

Lavery, D. Cockton, G., & Atkinson, M.P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, *16,* 246-266.

Lewis, J.R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors*, *36*, 368-378.

Lindgaard, G. (1999). Does emotional appeal determine the usability of web sites. *CYBERG '99, Western* Australia.

Lindgaard, G. (1994). *Usability testing and system evaluation: A guide for designing useful computer systems.* London: Chapman & Hall.

McClelland, I. (1998). Damaged merchandise: How might we fix it? *Human-Computer Interaction*, *13,* 283-288.

Olson, G.M., & Moran, T.P. (1998). Commentary on "Damaged merchandise?". *Human-Computer Interaction*, *13*, 263-323.

Nielsen, J. (1993). *Usability engineering*. Boston: Academic Press.

Nielsen, J., & Landauer, T. (1993). A mathematical model of the finding of usability problems. *Proceedings INTERCHI'93 Conference of Human Factors in Computing Systems* (pp. 206-213). New York: ACM Press.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings ACM CHI Conference on Human Factors in Computing Systems* (pp. 214-221). New York: ACM Press.

Sears, A. (1997). Heuristic walkthroughs: finding the problems without the noise. *International Journal of Human-Computer Interaction*, *9,* 213-234.

Swets, J.A. (1954). *Signal detection and recognition by human observers*. New York: John Wiley & Sons.